QUALITY CONTROL AND ANALYSIS OF ERROR IN THE NAVAL MANPOWER DATA BASE

James O'Neill Carter

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

QUALITY CONTROL AND ANALYSIS OF ERROR
IN
THE NAVAL MANPOWER DATA BASE

by

James O'Neill Carter

Thesis Advisor: K. T. Marshall

September 1972

Quality Control and Analysis of Error
in
The Naval Manpower Data Base

by

James O'Neill Carter
Lieutenant Commander, United States Navy
B.S., United States Naval Academy, 1963

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1972

ABSTRACT

This thesis develops a model which describes how errors enter and leave an operating data base for manpower management. The model describes the error input process and error distribution in the data base. The underlying structure for the model is the M/G/∞ queue. The model is used to determine the effect of a change in the input error rate on the number of errors in the data base. An upper limit is determined for this rate of increase, and a method of determining a minimum time between samples in the worst possible case is proposed.

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

Decision making processes in the Department of Defense are not unlike those in other large governmental and non-governmental industrial enterprises. During the past fifteen years, a considerable portion of the logistics, engineering, and management effort has been computerized. This has resulted in a considerable number of support and reference ADP files that constitute the data input for the computer. The files, or data bases, vary in size from 50,000 up to millions of records. In terms of alphanumeric characters, some of the files have from fifty million to ten billion characters. We will use the operating data base of the Navy's Bureau of Personnel as an example throughout this thesis, but the model developed is generally applicable to data bases in the logistics and engineering areas as well.

The Active Duty Enlisted Master Magnetic Tape Record (E.M.T.) is the operating data base which this thesis addresses. It contains 550 systematically arranged alphanumeric characters for every enlisted man on active duty, approximately 600,000 men. These alphanumeric characters represent such information as name, rate, serial number, social security number, age, race, religion, number of dependents, GCT/ARI scores, home of record, years of formal education, pay entry base date, duty station, and many others. For a detailed description of the contents of the data base, see Ref. [4]. This information is

used by manpower managers in the Bureau of Naval Personnel to facilitate assignments, to fill school quotas, to determine force parameters, to make budget and end strength predictions and many other manpower management decisions.

Inputs are made daily to the Enlisted Master Tape by every reporting unit in the Navy. This is done in the form of a unit diary. The diary is the paper that is submitted daily to an ADP center for editing, coding, and eventual insertion into the E.M.T. For example, see Fig. 1. Information flows from the reporting units to the ADP centers to the change routine which alters the E.M.T. See Ref. [2].

The purpose of this thesis is to develop a model which describes how errors enter and leave the data base (E.M.T.) and to investigate how this model can be used to help design sampling methods similar to those in standard statistical quality control procedures. It does not address format editing, which is covered in Ref. [2]. This function takes place at the ADP center and during the change routine. If the format is not correct for the type of data element being changed, the computer will not perform the change and so indicates. We are concerned in this thesis with an after-the-fact evaluation of the data in the E.M.T. We are concerned, then, with technical editing. Some examples of technical editing are these: correct service number, correct pay entry base date, correct rate, correct time in grade, and correct duty station. The results of the study will show that errors arrive in a Poisson fashion, that the number of errors in the base has a Poisson Distribution, and we show how this distribution changes

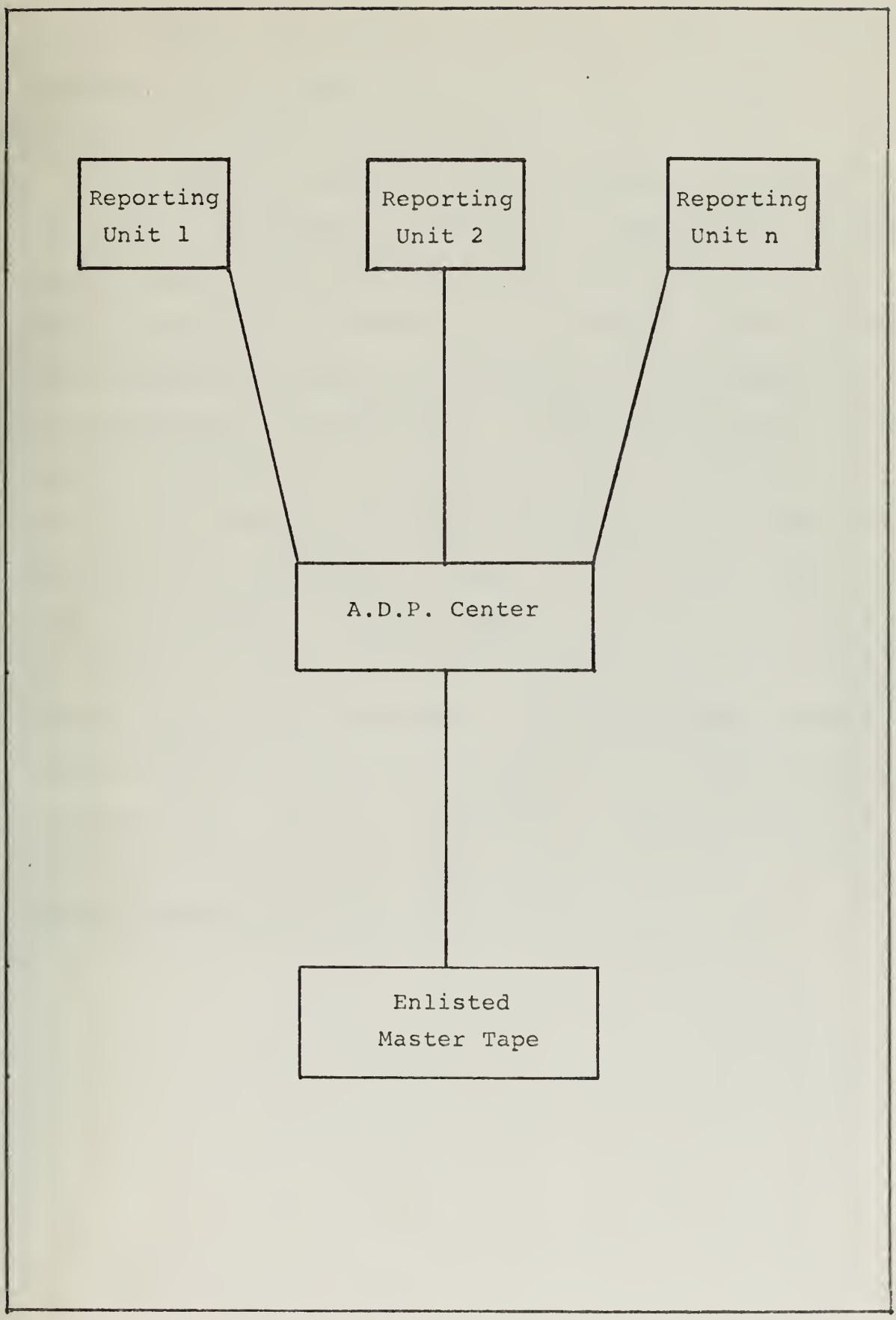

Figure 1. Information Flow from Reporting Units to the Enlisted Master Tape.

with a change in the error input rate. A sampling method is described and an upper limit for the rate of increase of errors in the system is determined.

This thesis is written in five sections, of which this is the first. In Section II we model the input process by which errors enter the data base. We show that they enter in a Poisson process. In Section III we develop a model for the distribution of errors in the data base. Our assumptions lead to formulating the $M/G/\infty$ queue. (The infinite server Poisson queue.) In Section IV we use this model to determine the effect of a change in the input error rate on the number of errors in the data base. An upper limit is determined for the rate of increase of the mean number of errors in the data base. In Section V we use the results of Section IV to help design error rate sampling procedures. The relationship between the size of the sample and the frequency of the sampling procedure is described. The goal is to design a sampling procedure in the data base which will allow an early detection of significant changes in the input error rate.

II. ARRIVAL OF ERRORS INTO THE DATA BASE

Assume a large number of possible places from which data can come each day, e.g.: 600,000 men, each with 50 data elements = 30×10^6 possible arrivals each day. Only about 1000 changes occur per day, so the probability any data element changes is about $1000/30 \times 10^6 \simeq .3 \times 10^{-4}$. A small proportion of these are in error, so the probability an error arrives is even smaller.

For each change to the data base which arrives, we define a Bernoulli random variable X_i. When change i is made, X_i takes on a value of zero if change i is correct and of one if change i is in error. When k changes are made, the number of errors which occur is a Binomial random variable if the following two conditions are satisfied: First, the probability of an error in any change is independent of other changes, Second, this probability is constant, i.e.: the X_i's are independent, identically distributed random variables.

It seems reasonable to assume that the receipt of an error from one reporting unit is independent of the receipt of an error from any other reporting unit. An exception might be a case in which an incorrect directive is being followed by all reporting units. This is the type of case which we would like to discover has occurred; however, for normal operations, we can assume it is not the case. We have no data to test whether or not the X_i's are identically distributed. This assumption implies, for example, that the probability an arriving data

element (such as pay entry base date) is in error is the same no matter where it came from or when it arrives.

The Poisson approximation to the Binomial is justified in cases where n is large and p is small. This is precisely the case under consideration. The number of possible changes, n in the data base is very large ($\simeq 30 \times 10^6$), and p, the probability of receipt of an error, is very small. Thus, in the remainder of the thesis, we assume that errors arrive in the data base in a Poisson process.

III. MODEL FOR THE DISTRIBUTION OF ERRORS

Assume that each day a number of errors arrive, this number being a Poisson random variable. Each one enters the data base and remains there until, at some future date, this data point is changed. It is changed by either of the following events:

a. a correct updated version of the data point arrives and replaces what is already there, or

b. an incorrect updated version of the data point arrives and replaces what is already there.

From our Poisson assumptions, we know that the events a. and b. are independent. Since the probability of replacing a data point which is currently in error with a new data point which is also in error is very small, we can assume that almost all the time, events of type a. are the only ones which remove errors from the data base.

Thus, the time an error spends in the data base is essentially independent of the arrival stream of errors.

Therefore the data base acts like the $M/G/\infty$ queue. (The infinite server queue, with Poisson arrivals, and a general service time distribution.)

Assume that the errors arrive at the data base at rate λ, that each one stays in the data base a random time, and that service time (length of stay in the data base) is randomly distributed as G.

Now, using the model of the infinite server Poisson queue, the problem of solving for the distribution of the number of errors in the data base is addressed. For the $M/G/\infty$ queue, with arrivals at rate λ and mean service time (average time a data point spends in the data base) of $1/\mu$, the number of errors in the system in steady state has a Poisson distribution with mean λ/μ [Ross Ref. 1, p. 18].

In general, let $X(t)$ denote the number of errors in the system at time t, where we start with no errors at time 0. We determine the distribution of $X(t)$ by conditioning on $N(t)$, the total number of errors which have arrived by time t. By conditioning, we obtain

$$p\{X(t)=j\} = \sum_{n=0}^{\infty} p\{X(t)=j \mid N(t)=n\} e^{-\lambda t} \frac{(\lambda t)^n}{n!} . \qquad (1)$$

The possibility that an error which arrives at time x will still be present at time t is $1-G(t-x)$. Hence, given that $N(t)=n$, it follows that the probability an arbitrary one of these errors is still present at time t is given by

$$p = \int_0^t (1-G(t-x))\frac{dx}{t} = \int_0^t (1-G(x))\frac{dx}{t} , \qquad (2)$$

independently of the others. This follows since we know that given $N(t)=n$, the n arrival times $S_1, \ldots, S_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval $(0,t)$ [Ross Ref. 1, p. 17].

Hence,

$$p\{X(t)=j \mid N(t)=n\} = \begin{cases} \binom{n}{j} p^j (1-p)^{n-j} & j=0,1,2,\ldots,n \\ 0 & j>n \end{cases} .$$

Thus, by (1) we have,

$$p\{X(t)=j\} = \sum_{n=j}^{\infty} \binom{n}{j} p^j (1-p)^{n-j} e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

$$= e^{-\lambda t} \frac{(\lambda tp)^j}{j!} \sum_{n=j}^{\infty} \frac{(\lambda t(1-p))^{n-j}}{(n-j)!}$$

$$= e^{-\lambda tp} \frac{(\lambda tp)^j}{j!} ,$$

where

$$p = \int_0^t \frac{(1-G(x))dx}{t} .$$

That is, X(t) has a Poisson distribution with mean =

$$\lambda \int_0^t (1-G(x))dx .$$

IV. CHANGE IN THE INPUT ERROR RATE

For the M/G/∞ queue with arrivals at rate λ and mean service time (average time a data point spends in the data base) of $1/\mu$, the number in the system (errors in the system) in steady state has a Poisson distribution with mean λ/μ [Ross Ref. 1, pp. 17, 18, 19].

Assume that for $t \leq t_1$, the data base errors have been arriving at a constant rate λ for some time and that the system is in steady state. At t_1, the error rate changes to a new rate β which for simplicity we assume is greater than λ. Thus in steady state at this new rate, the expected number of errors in the data base is β/μ. We wish to investigate how fast the expected number of errors reaches this level. (See Fig. 2.) Let $X(t)$ = the number of errors in the data base at time (t);

and $Y(t,x)$ = the number of errors in the data base at time $(t+x)$ that were in the data base at time (t), $x \geq 0$;

and $Z(t,x)$ = the number of errors in the data base at time $(t+x)$ that arrived during the interval $(t,t+x)$.

Then

$$X(t+x) = Y(t,x) + Z(t,x). \tag{3}$$

It is reasonable to assume that $Y(t,x)$ and $Z(t,x)$ are independent random variables. They are dependent to the extent that an incoming error might replace an error that is

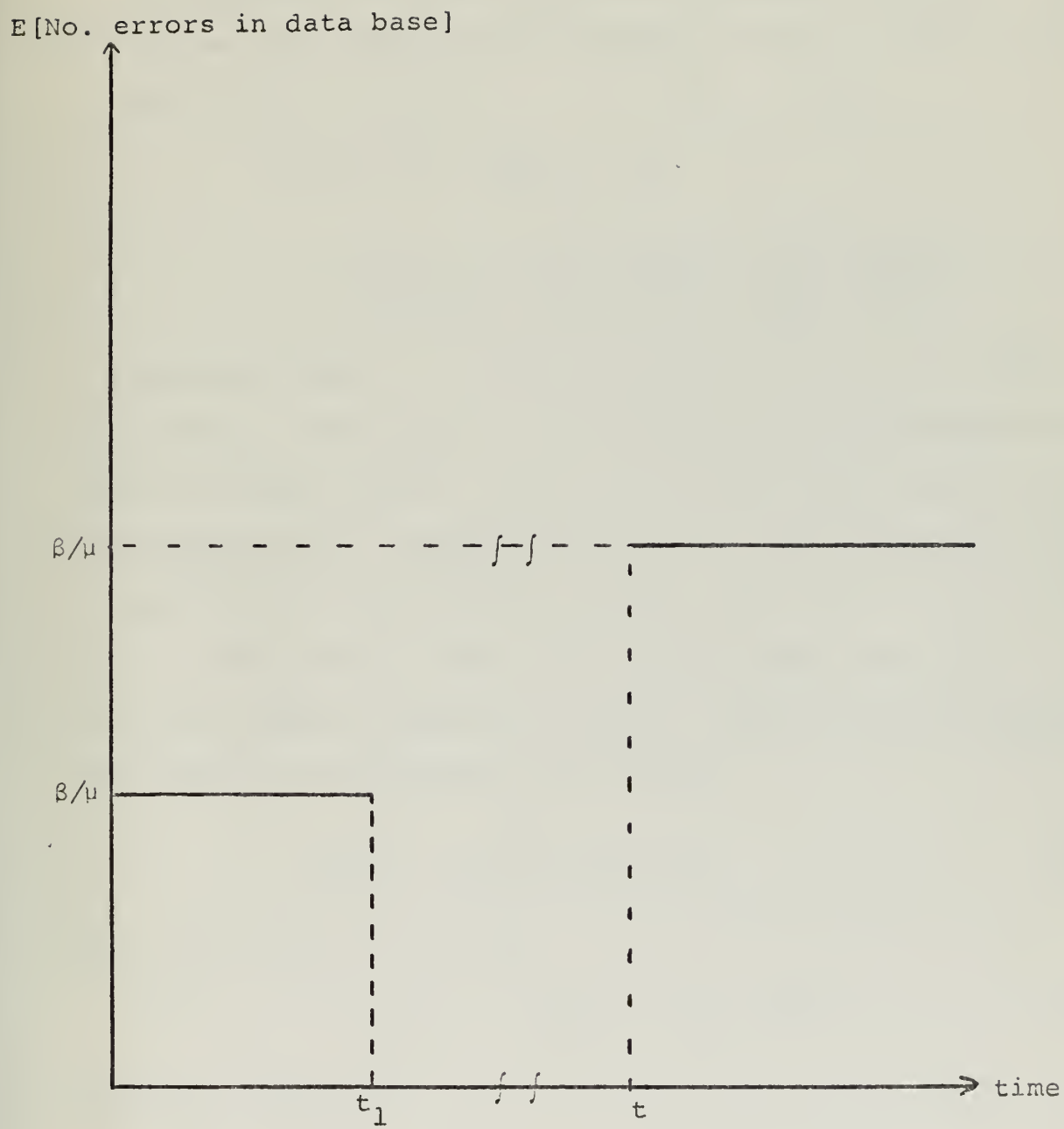

Figure 2. Expected Number of Errors in the Data Base vs. Time.

already there. We assume that such an event occurs only very rarely.

Our problem is to find the distribution of $X(t_1+x)$. To do this we must find the distribution of $Y(t_1,x)$ and $Z(t_1,x)$.

For $Y(t_1,x)$, we have

$$p\{Y(t_1,x) = k \mid X(t_1) = n\}$$

$$= p\{(n-k) \text{ of the errors in the system at time } t_1, \text{ have left by time } (t_1+x)\}.$$

An error which is in the system at time (t_1), has left the system by time (t_1+x) if and only if, its remaining service time is no more than x. If t_1 is an arbitrarily chosen point, then the remaining service time will be distributed the same as an equilibrium excess random variable [Ross Ref. 1, pp. 44-47]. That is, let service time (S) be distributed as G, and $E(S)=1/\mu$, then the remaining service time at t_1 for an arbitrary error is distributed $G_e(y)$ where

$$G_e(y) = \mu \int_0^y (1-G(u))du \ . \tag{4}$$

With this notation, we have

$$p\{Y(t_1,x) \mid X(t_1)=n\} = \binom{n}{k} \bar{G}_e(x)^k G_e(x)^{n-k} \ , \tag{5}$$

where $\bar{G}_e(x)=1-G_e(x)$.

We also know that [see Ross Ref. 1, pp. 18, 19]

$$p\{X(t_1)=n\} = \left(\frac{\lambda}{\mu}\right)^n \frac{e^{-\lambda/\mu}}{n!} \qquad n=0,1,2,\ldots \ . \tag{6}$$

By conditioning on $X(t_1)$ we have that

The inequality (10), when used in (9) shows that the true mean value function is bounded above by,

$$E(X(t_1+x)) \leq \lambda/\mu + (\beta-\lambda)x \quad \text{if } x \leq 1/\mu$$

$$\leq \beta/\mu \quad \text{if } x > 1/\mu .$$

This is shown in Fig. 3.

Figure 3 is useful in that it shows the maximum rate at which the mean number of errors in the data base adjusts to its new equilibrium value when the error input rate changes. This will give us an idea of how frequently to sample the data base, to see if error rates are changing. This we discussed in Section V.

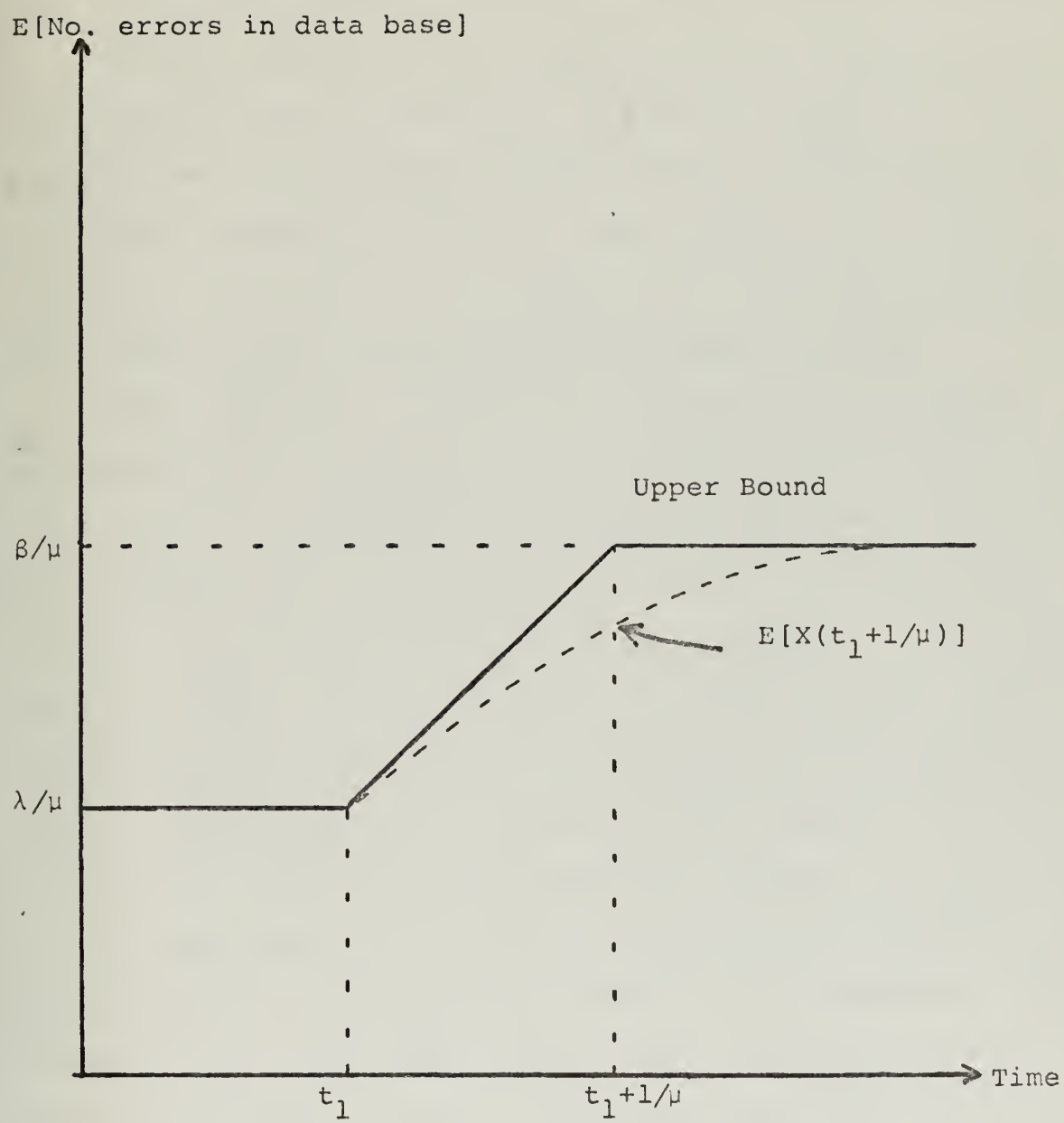

Figure 3. Upper Bound on the Mean of a Changing Input Error Rate.

V. SAMPLING PROCEDURE AND STATISTICAL QUALITY CONTROL

Quality control of, and sampling from, a population which has numerous input streams all subject to human error can be approached using standard quality control procedures if the following assumption is made: many separate reporting units following the same set of instructions with regard to diary and service record entries act as a single system.

This assumption is the basis for the quality control procedures now being used in a naval manpower data base, the U.S. Marine Corps Manpower Management System (see Ref. [2]). The method used is as follows: In order to determine the fraction of errors in the data base, a sample of 2500 (out of about 200,000) service records is randomly selected and compared with the source documents. The sampled data is sent, via U.S. mail, to the individual's reporting unit, under a cover letter asking for match/mismatch information between the information in the data base and the correct records at his reporting unit. Verification of the data is limited to the following: match, mismatch, or "can't find." The first indicates no error. The second indicates an error. The third indicates a case which arises when an individual is in transit between duty stations or not at his last known command. This last case could occur for many reasons, leave and temporary duty assignments elsewhere being the most likely. These cases are removed from the sample, and the fractional error rate for a given type of data

element is found by dividing the total mismatches by the sum of the mismatches and matches.

When the match/mismatch information is returned, the mean and variance of the fraction of errors for each element is determined as follows. Any process which generates output that can be characterized as either "correct" or "incorrect" for which each generalizing event (trial) is independent in the sense that it is not influenced by prior events and does not influence subsequent events and which can be described by a single parameter giving the probability of correct (or incorrect) events, is called a Bernoulli process. The probability of exactly c correct (or incorrect) events in n trials of such a process for the parameter p is given by the binomial distribution. For purposes of this study, the finite population of elements in the data base is so large that we have assumed the population to be infinite. This assumption is common in cases of acceptance sampling [Fetter Ref. 3, Chapter 1].

Let the sample size be n, and let the number of errors be a random variable X, where X is distributed Binomial (n,p). Then

$$\hat{p}=(\frac{X}{n}), \text{ so that } E(\hat{p}) = E(\frac{X}{n}) = \frac{1}{n} E(X) = \frac{1}{n}(np) = p .$$

Also,

$$Var(\hat{p}) = Var(\frac{X}{n}) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} npq = \frac{pq}{n} ,$$

Then

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} .$$

Now, we know that $E(X) = np$, and $Var(X) = npq$, thus

$$\sigma_x = \sqrt{npq} \therefore \frac{\sigma_x}{n} = \sqrt{\frac{pq}{n}} = \sigma_{\hat{p}} \tag{11}$$

To estimate the Variance of X from our single sample, define $\hat{\sigma}_x^2 = \hat{Var}(X)$. Then let

$$\hat{\sigma}_x^2 = n\hat{p}\hat{q} = n(\frac{X}{n})(1-\frac{X}{n}) = X - \frac{X^2}{n} . \tag{12}$$

Now we determine if this estimate is biased, and we find that

$$E[\hat{\sigma}_x^2] = E[X-\frac{X^2}{n}] = E[X] - E[\frac{X^2}{n}]$$

$$= np - \frac{1}{n} E[X^2] = np - \frac{1}{n} [np(1-p)+n^2p^2]$$

$$= (n-1)p(1-p),$$

since

$$E[X^2] = Var[X] + \{E[X]\}^2 = npq + n^2p^2 .$$

Thus we see that the above definition of $\hat{\sigma}_x^2$ is a biased estimate of σ_x^2. We can eliminate the bias by introducing the factor $(\frac{n}{n-1})$ into Eq. (12). Thus, we redefine

$$\hat{\sigma}_x^2 = \frac{n}{n-1} X(1 - \frac{X}{n}) . \tag{13}$$

It can be seen from the following calculations that this is an unbiased estimate of the variance.

$$E[\hat{\sigma}_x^2] = \frac{n}{n-1} \{E[X] - \frac{1}{n} E[X^2]\}$$

$$= \frac{n}{n-1} [np - \frac{1}{n}(np(1-p) + n^2p^2)]$$

$$= \frac{n}{n-1}[(n-1)p(1-p)] = npq .$$

With this estimate of the variance, the width of the 95% confidence interval can be determined as followed: for the sample of 2500 let X = 250.

Then

$$\hat{p} = \frac{250}{2500} = .1 .$$

Then by (13),

$$\hat{\sigma}_x^2 = \frac{n}{n-1}(X)(1 - \frac{X}{n}) = 225.$$

Thus

$$\hat{\sigma}_x = 15 \quad \text{and} \quad \hat{\sigma}_{\hat{p}} = \frac{15}{2500} = .006.$$

With this estimate of the standard deviation of the fraction of error (Fig. 4), we see that the 95% confidence interval is of width 2 x (1.96) x (.006) $\simeq$ .024, since for large n, binomial probabilities may be approximated by the normal distribution. The values of the parameters determined from this sample are taken to be the population parameters. When considering samples of size 200,000 (the whole population) and of size 200, with reference to the sample 2500, plotted in Fig. 4, it can be seen that the width of the confidence interval is inversely proportional to the sample size. For a sample which is equal to the whole population, the fraction of errors discovered would not be an estimate, but would in fact be the true fraction of errors. The width of the confidence interval would be zero, as shown by the line at .1 in Fig. 4.

For a sample of size 200, by Eq. (13) we would have:

$$\hat{\sigma}_x^2 = \frac{n}{n-1}(X)(1-\frac{X}{n}) = \frac{200}{199}(20)(.9) = 18 ,$$

thus

$$\hat{\sigma}_x = 4.25 \quad \text{and} \quad \hat{\sigma}_{\hat{p}} = \frac{4.25}{200} = .021.$$

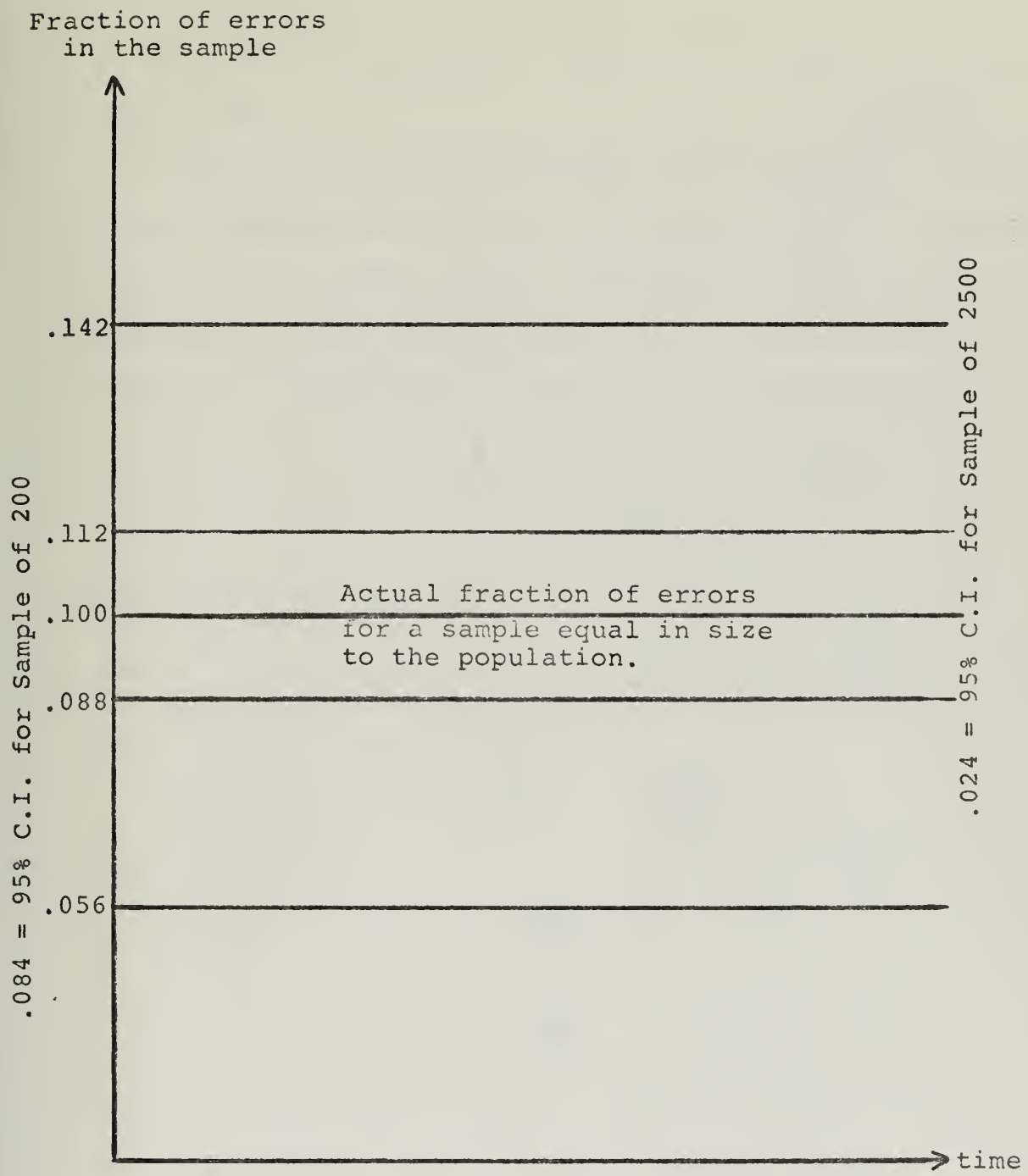

Figure 4. Comparison of Confidence Interval Width for Three Different Sample Sizes.

Then the 95% confidence interval is of width 2 x (1.96) x (.021) ≃ .084.

Sample size can be determined using the above procedure if we know the confidence level and interval width we desire and the population parameter (p). For example, if p=.1, and the half width of the interval, say α, = .02, and if we desire a 95% level of confidence in this interval, then we know that

$$1.96(\hat{\sigma}_{\hat{p}}) = .02$$

$$\hat{\sigma}_{\hat{p}} = \frac{.02}{1.96} \simeq .01$$

then by (11) we have that

$$.01 = \hat{\sigma}_{\hat{p}} = \frac{\hat{\sigma}_x}{n} \quad \text{or} \quad n = \frac{\hat{\sigma}_x}{.01}$$

thus

$$n = \frac{\hat{\sigma}_x}{.01} = \frac{\sqrt{npq}}{.01}$$

thus

$$n^2 = \frac{npq}{.0001}$$

$$n = \frac{pq}{.0001} = \frac{(.1)(.9)}{.0001} \simeq 900 .$$

See Figs. 5 and 6 for a plot of sample sizes vs. width of the confidence interval for p=.1 and .05 respectively.

Now having solved for the mean and variance of the fraction of errors in a sample of having determined the sample size which must be used in order to have a desired confidence interval, we are now able to address the question of frequency and sampling in terms of the minimum time between samples.

In order to be able to control the error rate, we must be able to predict its behavior. By the methods just described,

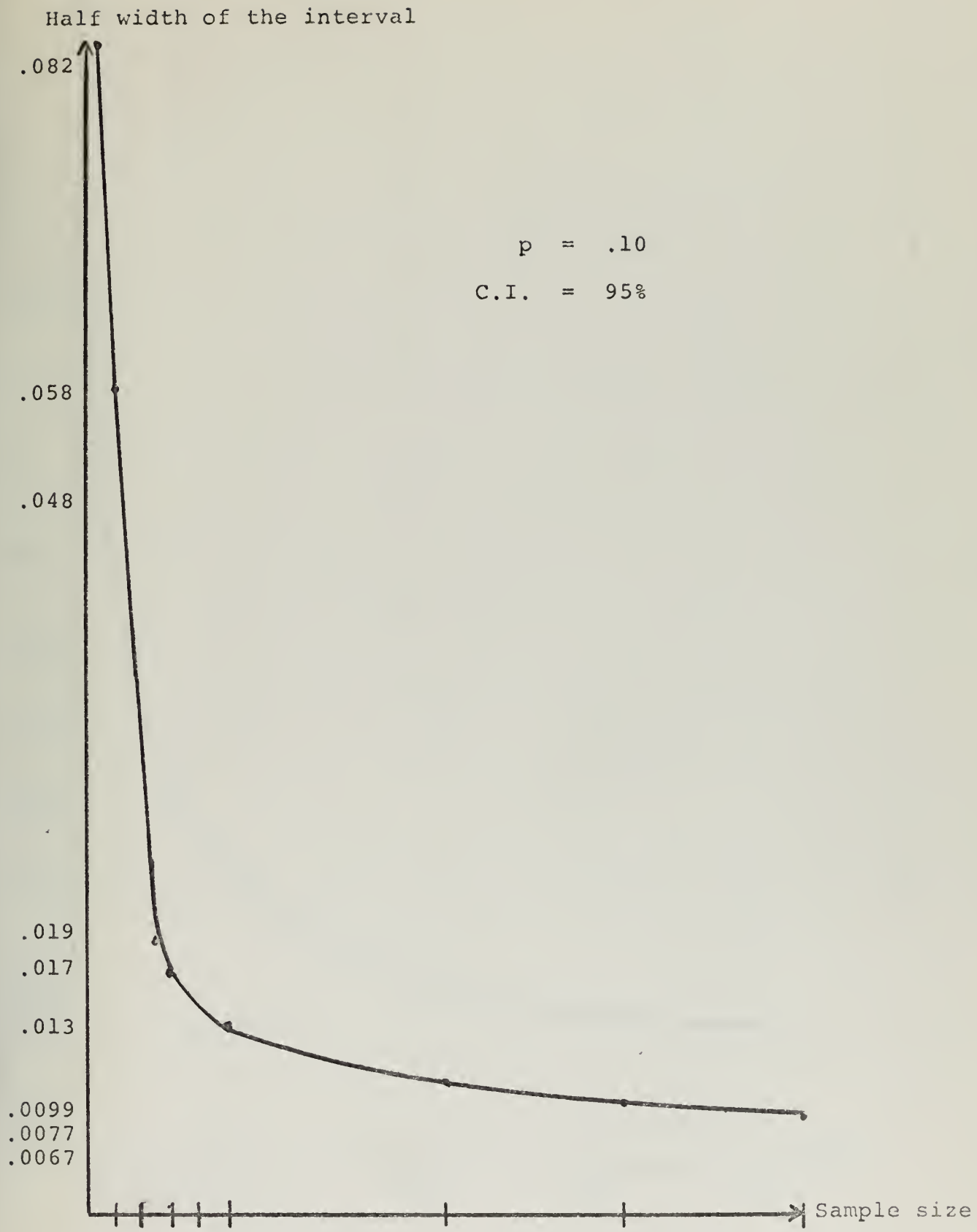

Figure 5. Plot os Sample Size vs. Confidence Interval Width for p=.1.

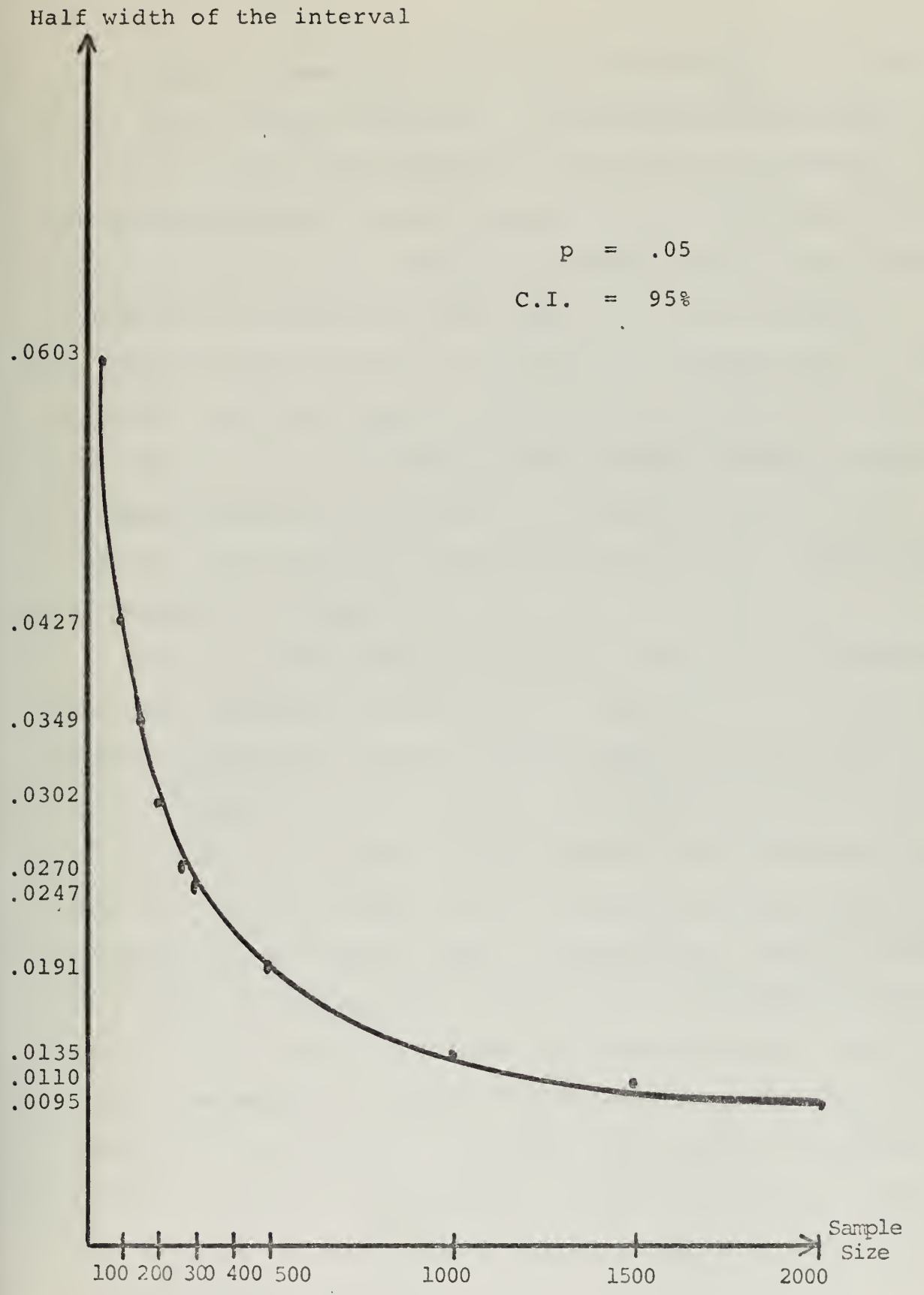

Figure 6. Plot of Sample Size vs. Confidence Interval Width for p=.05.

we are able to determine the limits within which the error rate from a sample should lie. By assuming that the same set of causal factors will continue to operate in the future, it is usually possible to make a prediction of the expected behavior of the system. Then, if a change occurs in the causal system which changes the error rate, this fact should be quickly apparent through an increase in the error rate in the samples. We then attempt to discover the cause of the increase and eliminate it. The time between samples becomes an important parameter when early detection of changes in error rate is desired. See the quality control chart in Fig. 7 [Fetter Ref. 3, Chapters 1 through 3].

As long as the fraction of errors stays within the upper and lower confidence bounds, we say that it is "in control." When it leaves this interval on the upper side, we say it is "out of control." If we were to pick limits of $\pm 1.96\ (\sigma)$ for our confidence interval, we would have a 95% confidence interval. That is, only five times out of a hundred would we think the process was "in control" when it was actually "out of control." If we were to increase the limits of our confidence interval to $\pm 3\sigma$, we would then have a 99+% confidence interval; that is, only three times out of a thousand would we mistakenly infer that the system was "in control." The wider the limits, the greater confidence we have in our interval. On the other hand, wider limits decrease the probability that a change in the process will be detected quickly.

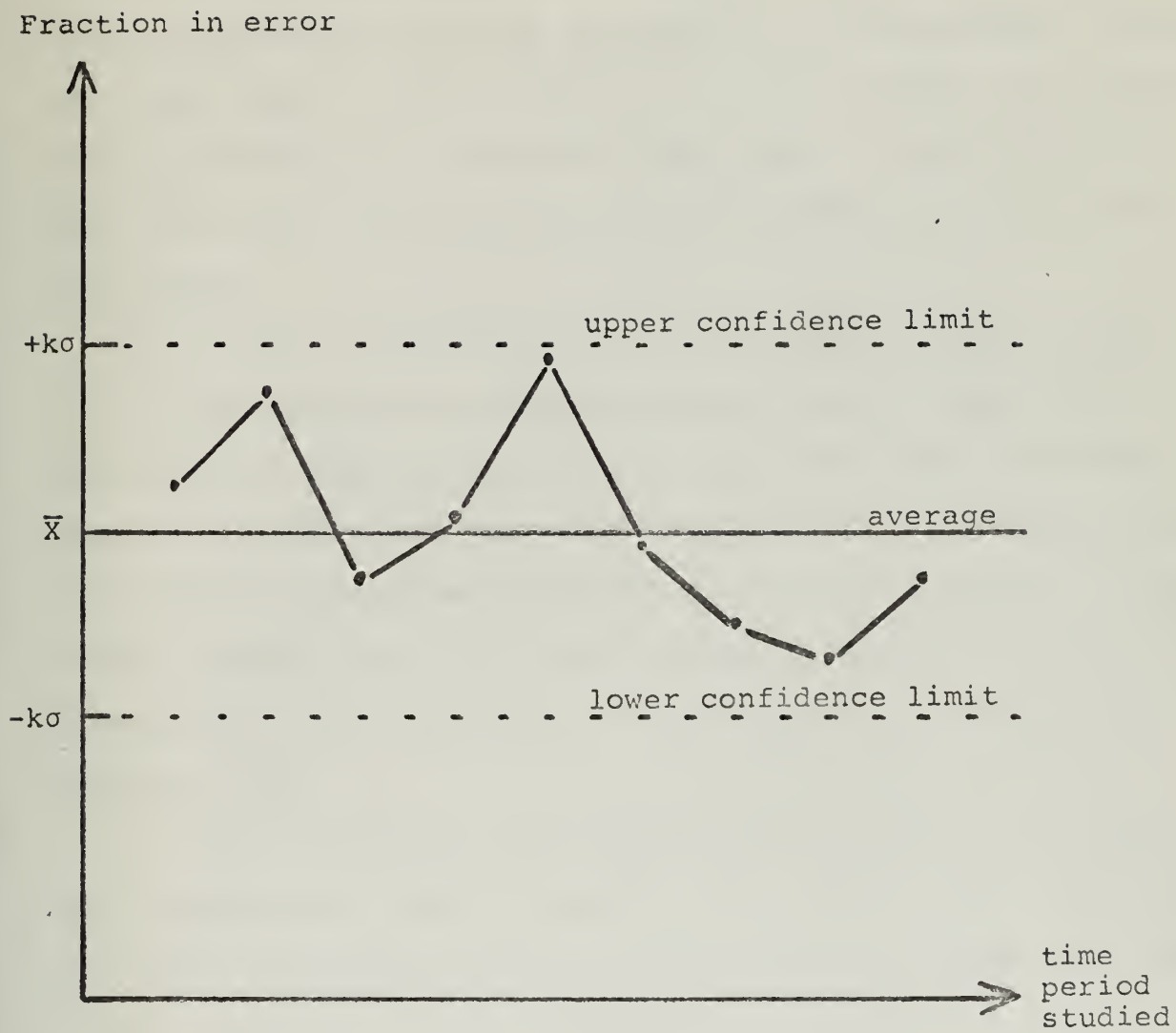

Figure 7. Typical Quality Control Chart.

Figure 8 shows changes in the average error from $\bar{X}_1$ to $\bar{X}_2$, with their corresponding normal densities. The shaded area represents the only place on the graph of the new density where an error rate will be determined to have changed, since any other place in the new density is still within the limits of the old density. The shaded area then is the probability that the shift will be detected by any sample taken following the change.

Recall Fig. 3 in Section IV for a change in input error rate. Now consider the worst possible case of input errors, when every change which arrives at the data base is an error. Then if at time t and thereafter every entry to the data base is an error, the number of errors in the data base will increase at its maximum rate. The upper bound on the fraction of errors present at time t, say p(t), is shown in Fig. 9 as the solid diagonal line.

We draw confidence limits about p(t) just as we do about the steady state lines. From Fig. 9 it can be seen that unless the lower confidence bound on p(t) exceeds the upper confidence bound on the p(0), we will probably not detect a change in the error rate.

The equation of p(t) is known since the average number of changes per day is known, and since all of these changes are assumed to be in error. We number the time scale so that the diagonal starts at t=0.

$$p(t) = p(0) + \mu(1-p(0))t; \quad 0<t\leq 1/\mu, \tag{14}$$

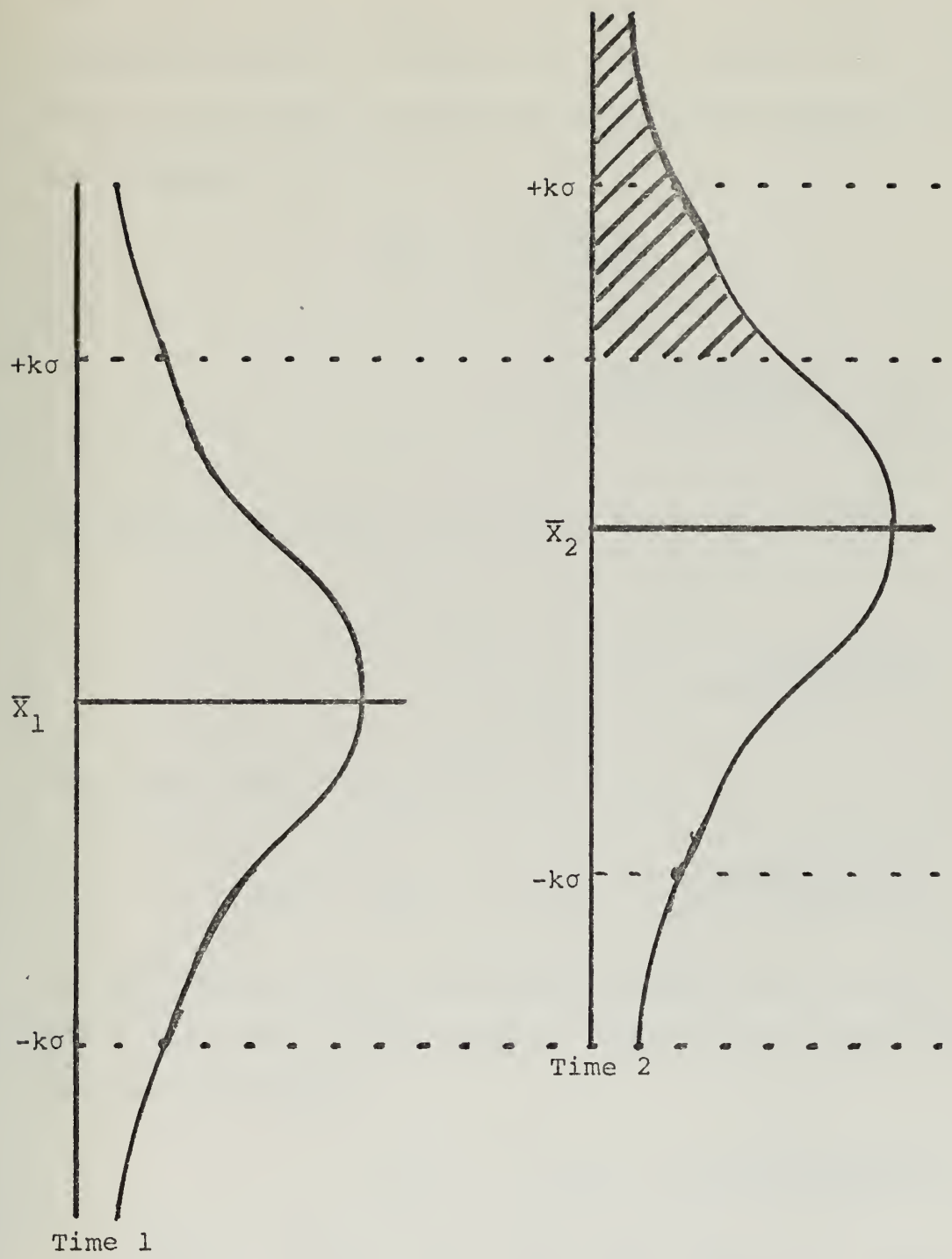

Figure 8. Change in Average Error and Normal Densities.

where $1/\mu$ is the time an error spends in the data base (assumed constant to obtain the upper bound in Fig. 9). Let $2\alpha(t)$ be the width of the 95% confidence interval at t. Then for a sample of size n,

$$\alpha(t) = \frac{1.96\sigma_x}{n}$$

$$= 1.96\sqrt{\frac{p(t)(1-p(t))}{n}} \quad , \quad 0<t\leq 1/\mu \ . \tag{15}$$

Let F(t) be the lower confidence limit at t. Then

$$F(t) = p(t) - \alpha(t)$$

$$= p(t) - 1.96\sqrt{\frac{p(t)(1-p(t))}{n}} \ . \tag{16}$$

The upper confidence limit at t=0 is

$$p(0) + \alpha(0) = p(0) + 1.96\sqrt{\frac{p(0)(1-p(0)}{n}} \ . \tag{17}$$

We take the minimum time between samples to be that t for which (16) and (17) are equal. That is, t_s, the time between samples, satisfies

$$F(t_s) = p(0) + 1.96\sqrt{\frac{p(0)(1-p(0))}{n}} \ . \tag{18}$$

For example, consider the case where p(0)=.1, a sample size n of 1000, a confidence interval of 95%, and a time in the data base of 300 days. Then

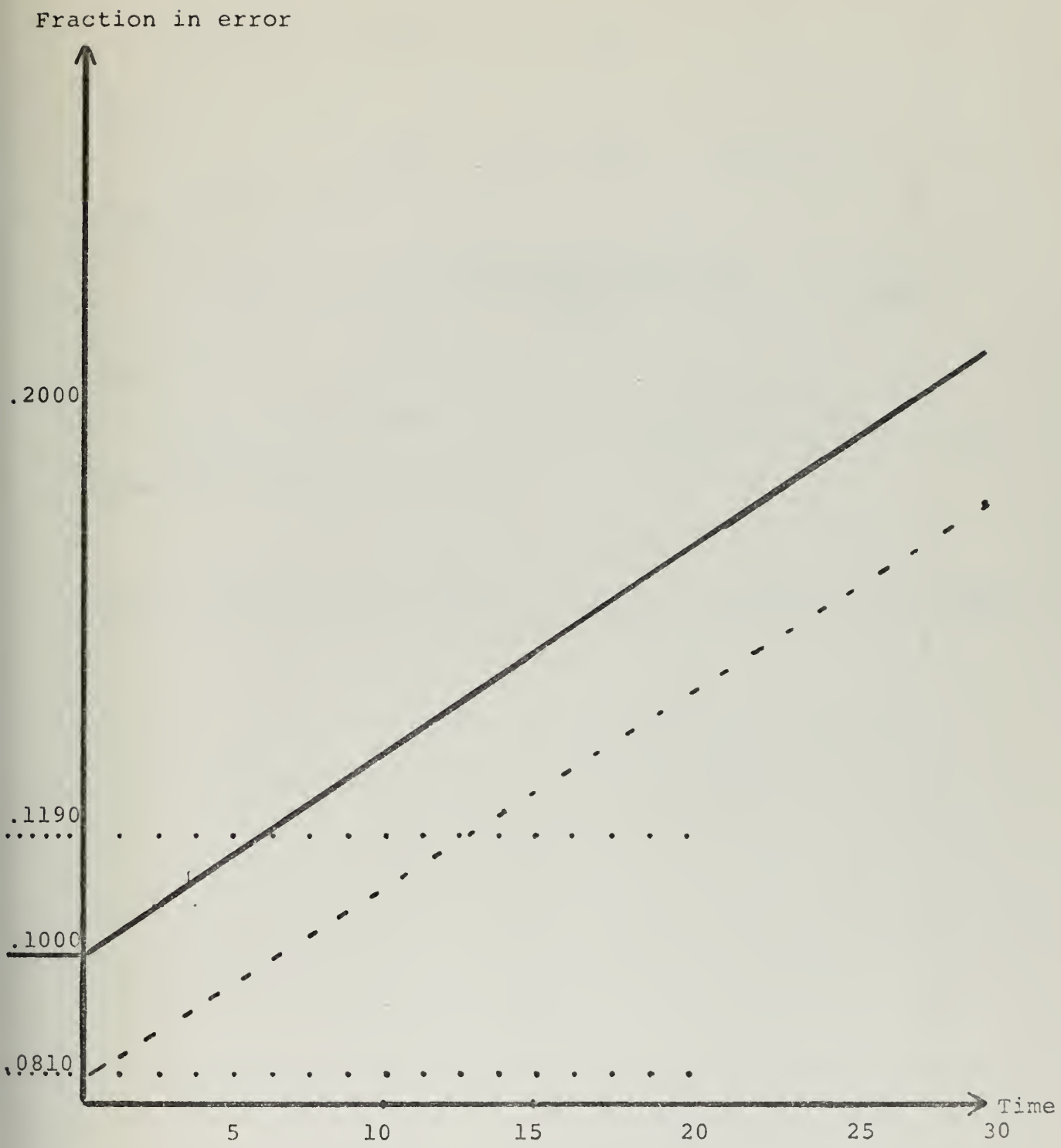

Figure 9. Confidence Interval for an Increasing Error Rate.

$$p(t) = .1 \ , \ t \leq 0$$

and

$$p(t) = 1 + .003t \ , \ 0<t<300 \ .$$

Also

$$\alpha(t) = 1.96 \sqrt{\frac{.09+.0024t-.000009t^2}{1000}} \ , \ 0 \leq t \leq 300 \ .$$

Thus

$$\alpha(0) = .019 \quad \text{and} \quad p(0) = .1 \ .$$

From Eq. (18) we find that

$$t_s \approx 14 \text{ days.}$$

That is, the minimum sampling period to consider is two weeks.

LIST OF REFERENCES

1. Ross, S. M., Applied Probability Models with Optimization Applications, Chapters 1-3, Holden-Day, 1970.

2. Office of Naval Research Report 145-e, Use of Edit Information to Measure and Reduce File Error Content a Mis, by R. W. Burton and S. C. Jaquette, October 1971.

3. Fetter, R. B., The Quality Control System, Chapters 1-3, Richard D. Irwin, Inc., 1967.

4. Bureau of Naval Personnel, Manual of the Active Duty Enlisted Master Magnetic Tape Record, NAVPERS 15,949C, July 1968.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Operations Research Program Office of Naval Research Washington, D.C. 20360	3
2.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
3.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
4.	Library, Code 55 Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
5.	Bureau of Naval Personnel (PERS-N) Washington, D.C. 20370	1
6.	CAPT H. V. Samis USN Bureau of Naval Personnel (PERS A12) Washington, D.C. 20370	1
7.	R. K. Lehto Bureau of Naval Personnel (PERS-AX) Washington, D.C. 20370	1
8.	LTCOL D. C. Pauley, USMC JUMPS/MMS Liaison Officer Marine Corps Finance Center Kansas City, Missouri 64194	1
9.	Commandant of the Marine Corps (AOIM) Headquarters Marine Corps Washington, D.C.	1
10.	Professor Kneale T. Marshall, Code 55 Mt Department of Operations Research and Administrative Science Naval Postgraduate School Monterey, California 93940	3

11.	Professor D. P. Gaver, Jr., Code 55 Gv Department of Operations Research and Administrative Science Naval Postgraduate School Monterey, California 93940	1
12.	Professor R. W. Butterworth, Code 55 Bd Department of Operations Research and Administrative Science Naval Postgraduate School Monterey, California 93940	1
13.	Professor P.A.W. Lewis, Code 55 Lw Department of Operations Research and Administrative Science Naval Postgraduate School Monterey, California 93940	1
14.	LCDR James O. Carter, USN 579 G Wilkes Lane Monterey, California 93940	1

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Naval Postgraduate School Monterey, California 93940	Unclassified
	2b. GROUP

3. REPORT TITLE

QUALITY CONTROL AND ANALYSIS OF ERROR
IN
THE NAVAL MANPOWER DATA BASE

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Master's Thesis, September 1972

5. AUTHOR(S) (First name, middle initial, last name)

James O'Neill Carter
Lieutenant Commander, United States Navy

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1972	39	4

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

This thesis develops a model which describes how errors enter and leave an operating data base for manpower management. The model describes the error input process and error distribution in the data base. The underlying structure for the model is the M/G/∞ queue. The model is used to determine the effect of a change in the input error rate on the number of errors in the data base. An upper limit is determined for this rate of increase, and a method of determining a minimum time between samples in the worst possible case is proposed.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

Unclassified
Security Classification

A-31408

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
error in data base						
manpower data base						
operating data base						
management information system						

138435

Thesis
C27372 Carter
c.1 Quality control and analysis of error in the Naval manpower base.